Working with Web Corpora

A practical/applied workshop

Organizers Felix Bildhauer <felix.bildhauer@fu-berlin>
Roland Schäfer <roland.schaefer@fu-berlin.de>
Freie Universität Berlin

Length 7 hours

While web corpora are often useful by virtue of being very large and containing unique patterns of linguistic variation, they cannot be used like traditionally compiled corpora in many respects. Prominent issues are the partially unknown composition of web corpora (where biases can arise due to the nature of the web or the data collection procedures), and the higher amount of noise (introduced at the source and by linguistic processing which is often not perfectly suited to the given kind of data). But even the extraordinary size itself can have implications on how data can be efficiently retrieved from web corpora, or how it should be analyzed statistically.

The first part of the workshop addresses a range of theoretical and practical issues arising in every-day linguistic research with web corpora, or with web data in general.

The main topics are:

- crawling methods and their impact on corpus composition
- duplication and deduplication methods
- handling of boilerplate
- normalization (e.g. hyphenation removal, orthographic correction)
- challenges in the linguistic annotation of web corpora
- classifying web documents (register/genre, topic, etc.)
- comparing web corpora to traditional corpora

In the second half of the workshop, we demonstrate how to work with web corpora in adequate and efficient ways, and we provide an introduction to the COW web corpora (available in Dutch, English, French, German, Spanish, Swedish) and the web-based Colibri^2 search tool (https://webcorpora.org), which users will use as part of the workshop for practical exercises.

This workshop aims primarily at linguists who are working (or are going to work) with web corpora and who wish to gain a better understanding of the peculiarities of their data and how to deal with them. The workshop will also benefit researchers with an interest in corpus compilation and the linguistic processing of web data.